Игорь С. Пантюхин, Никита К. Дружинин, Лев С. Титов,
Александр А. Капитонов, Алиса А. Воробьева
Университет ИТМО,
Кронверкский пр., 49, г. Санкт-Петербург, 197101, Россия
e-mail: zevall@ya.ru, http://orcid.org/0000-0002-3946-6057
e-mail: kleverteo@gmail.com, http://orcid.org/0000-0002-9898-2537
e-mail: letitov@yandex.ru, http://orcid.org/0000-0001-5246-3224
e-mail: kap2fox@gmail.com, http://orcid.org/0000-0001-5517-3038
e-mail: alice_w@mail.ru, http://orcid.org/0000-0001-6691-6167

СПОСОБ ИССЛЕДОВАНИЯ КОМПЬЮТЕРНЫХ ИНЦИДЕНТОВ НА ОСНОВЕ КЛАСТЕРИЗАЦИИ АТРИБУТОВ

DOI: http://dx.doi.org/10.26583/bit.2018.3.04

Аннотация. При проведении внутреннего аудита средств вычислительной техники важной задачей является снижение объемов хранимой и обрабатываемой информации. Требуется выделять группы схожих по некоторым параметрам информационных объектов и анализировать их отдельно. Подходящим инструментом для этого является оптимальная кластеризация данных. В данной работе представлен способ группирования файлов на жестком диске, основанный на алгоритме иерархической кластеризации Ланса - Вильямса. Файлы, относящиеся к одному компьютерному инциденту, будут находиться в одном и том же кластере. Данное утверждение основано на предположении, что на исследуемом устройстве пользователь совершал ряд связанных между собой действий по времени или по другому внешнему атрибуту или группе атрибутов (например, просмотр подряд некоторого числа изображений, составление и затем отправка электронного письма). В результате кластеризации эти данные сгруппируются в один кластер, в дальнейшем их можно будет представлять компьютерному криминалисту как потенциально совершенный компьютерный инцидент. Для апробации представленного способа был проведен эксперимент, в ходе которого были получены результаты, показывающие, что на тестовой системе файлы, созданные и просмотренные в один и тот же промежуток времени, оказались в одном кластере как при большом числе кластеров выходных данных способа, так и при малом.

Ключевые слова: кластеризация, внутренний аудит, компьютерная криминалистика, компьютерный инцидент, информационная безопасность.
Для цитирования: ПАНТЮХИН, Игорь С. et al. СПОСОБ ИССЛЕДОВАНИЯ КОМПЬЮТЕРНЫХ ИНЦИДЕНТОВ НА ОСНОВЕ КЛАСТЕРИЗАЦИИ АТРИБУТОВ. Безопасность информационных технологий, [S.l.], n. 3, p. 38-44, 2018. ISSN 2074-7136. Доступно на: <https://bit.mephi.ru/index.php/bit/article/view/1138>. Дата доступа: 28 aug. 2018. doi:http://dx.doi.org/10.26583/bit.2018.3.04.

Igor S. Pantiukhin, Nikita K. Druzhinin, Lev S. Titov,
Alexandr A. Kapitonov, Alisa A. Vorobeva
ITMO University,
Kronverkskiy pr., 49 Saint Petersburg, 197101, Russian Federation
e-mail: zevall@ya.ru, http://orcid.org/0000-0002-3946-6057
e-mail: kleverteo@gmail.com, http://orcid.org/0000-0002-9898-2537
e-mail: letitov@yandex.ru, http://orcid.org/0000-0001-5246-3224
e-mail: kap2fox@gmail.com, http://orcid.org/0000-0001-5517-3038
e-mail: alice_w@mail.ru, http://orcid.org/0000-0001-6691-6167

Method for investigating computer incidents based on attribute clustering

DOI: http://dx.doi.org/10.26583/bit.2018.3.04

Abstract. A reduction of the amount of stored and processed information is an important task for internal audit. It is required to select groups of informational objects with similar parameters and to analyze them separately. Optimal clustering of the data is a suitable method to solve this problem. This paper presents a method of files grouping on the hard disk, based on the Lance Williams algorithm of hierarchical clustering. Files with the same computer incident will belong to the same cluster. This statement is based on the assumption that the user has performed series of actions interrelated in time or in another external attribute or a group of attributes (for example, scanning a row of images in succession, compiling and

then sending an email) on the device under investigation. As a result of clustering, these data are grouped together into one cluster and further on they can be presented to a computer forensic scientist as a potential computer incident. Thus, there is no need to analyze the files itself, since the external file attributes such as creation time, access time, time of change, etc. are used as the meaningful parameters. This method also helps to specify the number of clusters manually for a rather flexible investigation of the tested file system. Experiment was carried on in order to test the presented method. The results of the experiment show that the files created and scanned within the same time interval ended up in the same cluster for both large and small number of the output data in the cluster.

Keywords: clustering, internal audit, computer forensics, computer incident, information security.
For citation: PANTIUKHIN, Igor S. et al. Method for investigating computer incidents based on attribute clustering. IT Security (Russia), [S.l.], n. 3, p. 38-44, 2018. ISSN 2074-7136. Available at: <https://bit.mephi.ru/index.php/bit/article/view/1138>. Date accessed: 28 aug. 2018. doi:http://dx.doi.org/10.26583/bit.2018.3.04.

Введение

В настоящее время из-за стремительного развития компьютерных технологий объемы хранимой и обрабатываемой информации растут из года в год. Появляются такие новые способы масштабируемого хранения неструктурированных данных, как data lake. Вместе с тем растет и число компьютерных преступлений.

Компьютерная криминалистика в последнее время сталкивается с целым рядом проблем из-за широкого использования самых разных видов информационных технологий. Они включают в себя растущий объем хранимых и обрабатываемых данных, увеличение числа типов устройств с разными платформами, каждое из которых необходимо анализировать отдельно, использование шифрования, а также появление новых парадигм информационных технологий, таких как облачные вычисления и Интернет вещей [1-3].

Применение интеллектуальных алгоритмов и способов позволяет достичь повышения точности и информативности получаемых сведений о компьютерном инциденте. Автоматизация инструментов для выполнения внутреннего аудита средств вычислительной техники существует преимущественно для выполнения простых задач, таких как поиск файлов, дешифровка данных и других [4]. Перспективность поиска новых способов автоматизации анализа, способных облегчить труд компьютерного криминалиста, вытекает из все большего роста числа компьютерных инцидентов в последнее время.

Для решения задачи повышения скорости исследования различных цифровых устройств в данной работе предложен способ кластеризации данных, в результате которого представляется возможность группировать вместе данные, потенциально представляющие собой сведения об инциденте на исследуемом средстве вычислительной техники [5]. Способ основан на иерархическом алгоритме кластеризации Ланса-Вильямса [6-7]. Он позволяет снизить объем хранимой и обрабатываемой информации (исследуются исключительно атрибуты файлов), а также позволяет исследовать компьютерные инциденты как на одном конкретном устройстве (компьютере, сервере, мобильном телефоне и других), так и на группе устройств.

Способ поиска инцидентов информационной безопасности с использованием иерархического алгоритма кластеризации

Подход основывается на предположении, что на исследуемом устройстве пользователь совершал не одно действие, а ряд связанных между собой действий по времени или по другому атрибуту (группе атрибутов). Например, просмотр подряд некоторого числа изображений в случае их неприемлемого содержания, составление и затем отправка электронного письма. Таким образом, отпадает необходимость анализировать сами файлы, так как в качестве значимых параметров используются числовые значения внешних файловых атрибутов, такие как время создания, время доступа, время изменения и другие. В результате кластеризации этих числовых значений файлы одного инцидента информационной безопасности сгруппируются в кластер,

который в дальнейшем можно представлять компьютерному криминалисту. Данный способ также позволяет вручную задавать число кластеров для более гибкого исследования тестируемой файловой системы.

Поскольку исследователь не знает конечное число кластеров данных, то предлагается использовать алгоритмы иерархической кластеризации. Они позволяют проводить кластеризацию при неизвестном числе кластеров и динамически показывать кластеры в зависимости от минимального расстояния между ними. Также это позволит исследователю вручную менять число кластеров в выходных данных для получения необходимого результата. Результатом такой кластеризации будет являться дендрограмма объединения данных в системе.

Для данной работы был выбран агломеративный алгоритм иерархической кластеризации Ланса-Вильямса, который представляет все значения данных в виде кластеров, число которых равно числу самих данных, а потом на основании расстояния между этими одноэлементными кластерами позволяет сгруппировывать их во все более и более крупные, пока наконец не останется один кластер, содержащий все элементы данной выборки [6-8].

Использование алгоритма Ланса - Вильямса для поиска инцидентов

Если объекты (или кластеры) i и j были объединены между собой в новый кластер $i \cup j$, то алгоритм Ланса - Вильямса позволяет вычислить расстояние между этим кластером и всеми оставшимися по рекурсивной формуле [9-10]:

$$d(i \cup j, k) = \alpha_i d(i,k) + \alpha_j d(j,k) + \beta d(i,j) + \gamma |d(i,k) - d(j,k)|, \quad (1)$$

где α_i, α_j, β, γ – агломеративные критерии.

Так как объекты кластеров представляются в виде точек в евклидовом пространстве, в качестве функции расстояния между элементами было выбрано Евклидово расстояние, позволяющее уменьшить влияние на конечный результат удаленных друг от друга объектов, а расстояние между кластерами рассчитывалось при помощи метода Варда, использующего подходы дисперсионного анализа для оценки расстояний между кластерами [10-12]. Способ минимизирует сумму квадратов для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге, то есть сначала в обоих кластерах для всех имеющихся наблюдений производится расчёт средних значений отдельных переменных. Затем вычисляются квадраты евклидовых расстояний от отдельных наблюдений каждого кластера до этого кластерного среднего значения. Эти дистанции суммируются. Потом в один новый кластер объединяются те кластеры при объединении которых получается наименьший прирост общей суммы дистанций.

Для метода Варда агломеративные критерии будут:

$$\alpha_i = \frac{|i| + |k|}{|i| + |j| + |k|}$$

$$\alpha_j = \frac{|j| + |k|}{|i| + |j| + |k|}$$

$$\beta = -\frac{|k|}{|i| + |j| + |k|}$$

$$\gamma = 0.$$

Координаты центра кластера, который объединяет в себе кластеры i и j, будет рассчитываться по формуле

$$g = \frac{|i| g_i + |j| g_j}{|i| + |j|}, \quad (2)$$

а расстояние между центрами кластеров g_i и g_j:

$$\frac{|i||j|}{|i|+|j|}||g_i - g_j||^2. \quad (3)$$

где $|i|$ – количество объектов в кластере i, а ||.|| -- евклидово расстояние

Вычислительный эксперимент

В ходе работы была составлена обучающая выборка, представляющая собой файлы с виртуальной машины Virtual Box под управлением операционной системы Windows XP, на которой была сформирована папка с изображениями с именами vocaloid_pack (x).jpg, где x – номер файла в тестовой выборке (всего 700 файлов). Затем данные изображения были открыты при помощи стандартных средств просмотра изображений Windows, чтобы сымитировать последовательный просмотр файлов пользователем. В качестве исходных данных были предложены числовые значения таких атрибутов файлов, как время последнего доступа, время модификации, время создания и расширение. В ходе кластеризации были проанализированы все файлы с исследуемой операционной системы в количестве 16 481 файл.

При помощи библиотеки scikit-learn для языка Python в ходе работы была получена дендрограмма, представляющая собой визуализацию в форме дерева, показывающая порядок и расстояние между объединениями в иерархической кластеризации (рис. 1).

При значительном увеличении, показанном на рис. 2 и на рис. 3, можно видеть, как объединились в кластер файлы с именами vocaloid_pack вне зависимости от их расширения.

Динамически меняя число кластеров, которое можно получить из данных полученной дендрограммы (от 16 481, когда кластеры содержат только по одному файлу, до 1, когда есть только один кластер, содержащий все исследуемые файлы), можно видеть, что при числе 65 кластеров тестовые изображения входят в один кластер, не включающий в себя другие файлы.

На рис. 4 представлен фрагмент вывода тестовой программы, в котором перечислены некоторые кластеры и представлен список файлов в каждом кластере, и показано, как в кластере под номером 6 сгруппировались исследуемые изображения. Также можно видеть, как в одной группе оказались файлы безопасности ОС Windows SAM и SECURITY, а в другой группе вместе оказались файлы, относящиеся к Virtual Box.

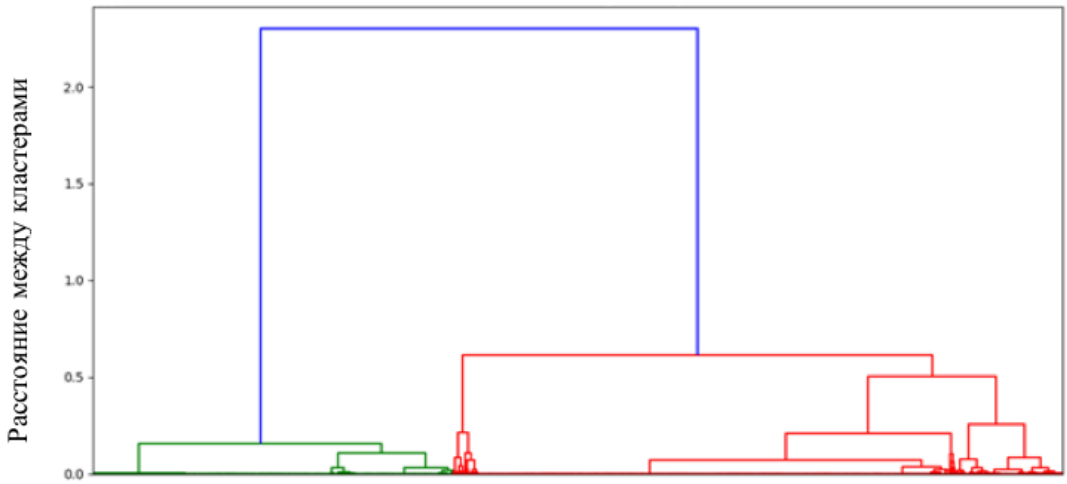

Рис. 1. Дендрограмма файлов исследуемой системы
(Fig. 1. Dendrogram of test system files)

Рис. 2. Дендрограмма при увеличении
(Fig. 2. Dendrogram with zooming)

Рис. 3. Дендрограмма при увеличении
(Fig. 3. Dendrogram with zooming)

```
0 manifest.json

1 A0000162.com

2 fp4amsft.dll fp4awel.dll fpmmc.dll

3 xpmsgr.chm lvback.gif type.wav

4 NTUSER.DAT NTUSER.DAT SAM default SECURITY

5 index.dat index.dat index.dat index.dat index.dat index.dat

6 vocaloid_pack (37).png vocaloid_pack (16).jpg vocaloid_pack (82).jpg vocaloid_pack
(119).jpg vocaloid_pack (143).jpg vocaloid_pack (180).jpg vocaloid_pack (166).jpg
vocaloid_pack (3).png vocaloid_pack (10).jpg vocaloid_pack (138).jpg vocaloid_pack (117).jpg
vocaloid_pack (198).jpg vocaloid_pack (223).jpg vocaloid_pack (45).png vocaloid_pack
(14).jpg vocaloid_pack (247).jpg vocaloid_pack (242).jpg vocaloid_pack (222).jpg
vocaloid_pack (211).jpg vocaloid_pack (142).jpg vocaloid_pack (162).jpg vocaloid_pack
(50).jpg vocaloid_pack (63).jpg vocaloid_pack (81).jpg vocaloid_pack (101).jpg vocaloid_pack
(15).png vocaloid_pack (11).png vocaloid_pack (25).png vocaloid_pack (38).jpg vocaloid_pack
(52).jpg vocaloid_pack (97).jpg vocaloid_pack.jpg vocaloid_pack (157).jpg vocaloid_pack
(178).jpg vocaloid_pack (245).jpg vocaloid_pack (241).jpg vocaloid_pack (215).jpg
vocaloid_pack (205).jpg vocaloid_pack (206).jpg vocaloid_pack (44).jpg vocaloid_pack
(55).jpg vocaloid_pack (120).jpg vocaloid_pack (61).jpg <...

7 History Temporary Internet Files A0000153.hhk A0000154.hhk A0000155.hhk A0000161.hhk
A0000160.hhk A0000156.hhk A0000159.hhk A0000157.hhk A0000158.hhk ntuser.ini ntuser.ini

8 A0000052.sys A0000121.dll A0000117.dll A0000112.dll A0000113.dll A0000094.exe A0000108.exe
moricons.dll batt.dll bthci.dll mdminst.dll sdhcinst.dll sti_ci.dll

9 eula.txt eula.txt eula.txt eula.txt eula.txt eula.txt eula.txt

10 empty.cat empty.cat empty.cat empty.cat empty.cat empty.cat empty.cat
11 VBoxMouse.sys VBoxVideo.sys VBoxVideo.sys VBoxGuest.sys VBoxGuest.sys oem3.inf oem4.inf
VBoxVideo.cat VBoxVideo.inf VBoxVideo.cat VBoxGuest.cat VBoxGuest.cat VBoxMouse.cat
VBoxVideo.inf VBoxGuest.inf VBoxMouse.inf VBoxGuest.inf oem2.inf VBoxDrvInst.exe
```

Рис. 4. Вывод тестовой программы
(Fig. 4. Test program output)

Заключение

Предлагаемый способ позволяет исследовать компьютерные инциденты с постинцидентных средств вычислительной техники. Использование алгоритма иерархической кластеризации, который в отличие от графовых и статистических алгоритмов выявляет детальную кластерную структуру множества объектов в виде дендрограммы, позволяет исследователю объединять в группу файлы компьютерных систем, относящиеся к одному инциденту информационной безопасности, а также проводить более гибкий анализ файловой системы, вручную задавая число кластеров. Представленный способ позволяет исследовать компьютерные инциденты анализируя только числовые значения атрибутов файлов, а не сами файлы, что снижает вычислительную сложность обработки данных. Таким образом, появляется возможность исследования компьютерных инцидентов в условиях постоянного роста объема хранимых и обрабатываемых данных. Апробация способа, представленная в работе, показала возможность его применения в системах исследования инцидентов информационной безопасности. Предлагаемый в работе способ универсален и может применяться для исследования компьютерных инцидентов в различных средствах вычислительной техники, таких как персональные компьютеры, серверы, мобильные устройства, дата-центры. Исследование компьютерных инцидентов, основанное на анализе атрибутов и их значений, имеет практическое применение [13] и может быть использовано в будущем при разработке предиктивных систем защиты от компьютерных инцидентов, для сокращения временных затрат при исследовании компьютерных инцидентов в больших объемах данных и иных задач в области компьютерной криминалистики [14].

СПИСОК ЛИТЕРАТУРЫ:

1. Mohay, G., Anderson, A., Collie, B., De Vel, O., McKemmish, R., Computer and intrusion forensics, Boston: Artech House, 2003
2. Caloyannides, M., Privacy protection and computer forensics, Boston: Artech House, 2001
3. Nelson B., Phillips A., Steuart C., Guide to Computer Forensics and Investigations, Boston: Cengage Learning, 2015
4. Пантюхин И.С., Зикратов И.А. Методика проведения постинцидентного внутреннего аудита средств вычислительной техники. Научно-технический вестник информационных технологий, механики и оптики. 2017. Т. 17. № 3. С.467–474.
5. Anderberg MR Cluster Analysis for Applications. Academic Press, New York, 1973.
6. Gavriel Yarmish, Philip Listowsky, Simon Dexter Distributed Lance-William Clustering Algorithm. 2017
7. Garg, A., Mangla, A., Gupta, N., & Bhatnagar, V. PBIRCH: A scalable parallel clustering algorithm for incremental data. In Database Engineering and Applications Symposium, 2006. IDEAS'06. 10th International (pp. 315–316).
8. Olson C.F. Parallel Algorithms for Hierarchical Clustering, Computer Science Division University of California at Berkeley Berkeley, 1993
9. Murtagh F., Contreras P. Methods of Hierarchical Clustering, 2011
10. Contreras P., Murtagh F. Fast hierarchical clustering from the Baire distance. In Classification as a Tool for Research, eds. H. Hocarek-Junge and C. Weihs, Springer, Berlin, 235–243, 2010.
11. Day W. H. E., Edelsbrunner H. Efficient algorithms for agglomerative hierarchical clustering methods Journal of Classification, 1984, 1: pp. 7–24
12. Gan G, Ma C., Wu J. Data Clustering Theory, Algorithms, and Applications Society for Industrial and Applied Mathematics. SIAM, 2007.
13. I. Pantiukhin, I. Zikratov, A. Sizykh and A. C. C. Nii, "Testing of the hypothesis in the research of computer incidents on the basis of the analysis of attributes and their values," 2017 20th Conference of Open Innovations Association (FRUCT), St. Petersburg, 2017, pp. 352-357. doi: 10.23919/FRUCT.2017.8071333
14. Пантюхин, И.С. Основы компьютерно-технической экспертизы / И.С. Пантюхин, Д.Н. Шидакова // Вестник полиции. - 2016. - № 1(7). - С. 20-29

REFERENCES:

[1] Mohay, G., Anderson, A., Collie, B., De Vel, O., McKemmish, R., Computer and intrusion forensics, Boston: Artech House, 2003
[2] Caloyannides, M., Privacy protection and computer forensics, Boston: Artech House, 2001
[3] Nelson B., Phillips A., Steuart C., Guide to Computer Forensics and Investigations, Boston: Cengage Learning, 2015
[4] Pantiukhin I.S., Zikratov I.A. Post-incident internal audit procedure of computer devices. Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2017, vol. 17, no. 3, pp. 467–474. (in Russian).
[5] Anderberg MR Cluster Analysis for Applications. Academic Press, New York, 1973.
[6] Gavriel Yarmish, Philip Listowsky, Simon Dexter Distributed Lance-William Clustering Algorithm. 2017
[7] Garg, A., Mangla, A., Gupta, N., & Bhatnagar, V. PBIRCH: A scalable parallel clustering algorithm for incremental data. In Database Engineering and Applications Symposium, 2006. IDEAS'06. 10th International (pp. 315–316).
[8] Olson C.F. Parallel Algorithms for Hierarchical Clustering, Computer Science Division University of California at Berkeley Berkeley, 1993
[9] Murtagh F., Contreras P. Methods of Hierarchical Clustering, 2011
[10] Contreras P., Murtagh F. Fast hierarchical clustering from the Baire distance. In Classification as a Tool for Research, eds. H. Hocarek-Junge and C. Weihs, Springer, Berlin, 235–243, 2010.
[11] Day W. H. E., Edelsbrunner H. Efficient algorithms for agglomerative hierarchical clustering methods Journal of Classification, 1984, 1: pp. 7–24
[12] Gan G, Ma C., Wu J. Data Clustering Theory, Algorithms, and Applications Society for Industrial and Applied Mathematics. SIAM, 2007.
[13] I. Pantiukhin, I. Zikratov, A. Sizykh and A. C. C. Nii, "Testing of the hypothesis in the research of computer incidents on the basis of the analysis of attributes and their values," 2017 20th Conference of Open Innovations Association (FRUCT), St. Petersburg, 2017, pp. 352-357. doi: 10.23919/FRUCT.2017.8071333
[14] Pantiukhin, Igor S., and Diana N. Shidakova. "Basics of Computer Forensics." Vestnik policii 1 (2016): 20-29. (in Russian).

Поступила в редакцию – 23 апреля 2018 г. Окончательный вариант – 23 августа 2018 г.
Received – April 23, 2018. The final version – August 23, 2018.